Association for Information Systems
AIS Electronic Library (AISeL)

AMCIS 2001 Proceedings

Americas Conference on Information Systems (AMCIS)

December 2001

Information Retrieval with Multiple Queries

Yunjie Xu
Syracuse University

INFORMATION RETRIEVAL WITH MULTIPLE QUERIES

Yunjie Xu
Syracuse University
yuxu@syr.edu

Abstract

Information overload becomes an immediate issue as the Internet prospers. To improve information retrieval effectiveness, we propose a sum-cosine procedure which is a meta-search procedure that integrates retrieval results from multiple queries. Based on this procedure, we also give its theoretical justification. We show that it improves performance through better estimation of user's information need and a solid integration method. We also show that meta-search with multiple queries is essentially a special form of relevance feedback. Our empirical evidence is fully consistent with theoretical analysis.

Keywords: Information retrieval, data fusion, meta-search, relevance feedback, multiple queries

Introduction

It is already a cliché to say the Internet age is an age of information overload. Information retrieval (IR) system such as search engine is usually an answer to such need. Unfortunately, search engines are far from satisfactory. Improving information retrieval effectiveness is and immediate need for both information consumer and information provider.

Since 1960's, many efforts had been put into this area. If we view the IR system as an integrated system of user, information, and search mechanism, the improvement of IR effectiveness can be achieved through any one of them (figure 1).

Research in information representation tries to impose certain structure on documents. The information representations explored by researchers include bag-of-words, XML, ontology etc. User behavior study tries to investigate how user expresses his intention and the change of his intention (Efthimiadis, 2000). User behavior study has strong implications on the design of automated web agent. But most researches are devoted to improving search mechanism. The objective here is to better match retrieved documents and user's information need. Different search strategies have been used to find what user wants. Vector space model is the earliest one. Bayesian classification, inference network, clustering based search, meta-search (also known as data fusion in IR) (Ng and Kantor 2000), genetic algorithm, classification tree, and neural network have all been tried recently.

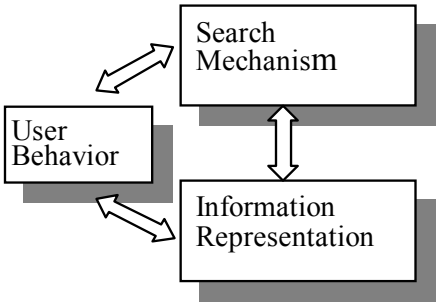


Figure 1. Major Streams of Research in IR

Our research could be viewed as one to improve search mechanism through meta-search. Meta-search technique tries to integrate search results from different search engines. In our research, we integrate search results from multiple queries that represent the same information need. These queries are generated automatically and the retrieval results for different query runs are integrated. The intuition is that multiple queries may better estimate or cover more aspects of user's information need. We propose a new procedure. Based on it, we investigate 1) Why and when such integration improve performance? 2) How to generate multiple queries? 3) How to integrate retrieval results? 4) The relationship between this method and relevance feedback that is commonly used in IR. Section 2 reviews the literature. In section 3 we discuss the procedure. Theoretical analysis of the procedure is given is section 4. Experimental results are given in section 5. Section 6 concludes.

Literature Review

Given an information representation method, because search mechanism is to match user's information need and documents, it can be improved through both better estimation of user's information need and better search strategy. There are two sub-areas of search mechanism that are closely related to our research and are reviewed in this section.

Meta-search

Meta-search in IR studies the integration of search results from multiple search schemes (A search scheme is a fixed combination of document representation method and search algorithm) (Ng and Kantor 2000). There are both successes and failures in data fusion research. It is still not clear why and when data fusion can increase recall and/or precision. Researches in this area tend to look for *ad hoc* methods that outperform benchmark (Belkin *et al* 1993, Bartell, *et al 1994*, Fox and Shaw 1994, Lee 1997). No rigorous justification of these methods was given.

Meta search with multiple queries can be traced back to (Katzer *et al.* 1982) who found that when user uses difference query language, the documents found have little overlap given same information need statement. Fox and Shaw (1994) found that combination of retrieval results from multiple dissimilar queries can improve performance. Belkin et al. (1993) show that combining multiple Boolean queries before it is submitted can improve performance.

Previous research in this area lacks theoretical justification. We propose the first fully automatic method and give theoretical and empirical justification.

Relevance Feedback

Our method resembles most to relevant feedback in IR. Relevance feedback technique asks user to classify initial retrieved result and use this information to revise initial query. The apparent difference between meta-search with multiple queries and relevance feedback is that relevance feedback revises query before final retrieval, while meta-search with multiple queries combines results after all retrieval runs. Our study shows that theoretically meta-search with multiple queries can be viewed as a special form of relevance feedback.

Relevance feedback is essentially a technique to re-estimate user's information need through feedback. The ultimate purpose of Rocchio's relevance feedback is to estimate an optimal query which is:

$$Q_{opt} = \frac{1}{n}\sum_{R}\frac{D_i}{|D_i|} - \frac{1}{N-n}\sum_{IR}\frac{D_i}{|D_i|}$$

where N is number of documents in the whole collection and n is number of relevant documents. Since n is not available beforehand, an adaptive term weighting methods is usually adopted to revise query (Rocchio 1965). This method can be described by the equation below.

$$Q' = \alpha Q + \beta\sum_{R}\frac{D_i}{|D_i|} - \gamma\sum_{IR}\frac{D_i}{|D_i|}$$

Where Q' and Q are the revised and the original query represented by word vectors, and α, β, and γ are coefficients less than 1. The difficulty of such adaptive learning is the choice of coefficients. Since there is no way to decide which set of value will lead to best performance, coefficients are set based on try and error.

The same difficulty exists in automatic relevance feedback in which top a few documents returned by retrieval system are treated as relevant automatically without user's proofread.

Buckley and Salton (1995) proposed a feedback formula that is better than Rocchio's. They use $\beta' 1/\|D_R\|$ and $\gamma' 1/\|D_{IR}\|$ as coefficient, where $\|D_R\|$ and $\|D_{IR}\|$ are size of relevant and irrelevant documents respectively. This method does not point out a way to set β, but it does try to mitigate the bias introduced by the number of document in relevant and irrelevant samples.

Our research shows that data fusion with multiple queries is essentially a special form of automatic relevance feedback. It is special in the sense that a sample of relevant documents is used to estimate optimal query, rather than using adaptive learning. Whether we combine queries or combine result does not affect the final result in our setup.

Sum-cosine Procedure

To investigate why data fusion with multiple queries can improve IR effectiveness, we first propose our procedure and then base our investigation on it. Our procedure is outlined as follows:

1. Use user's original query Q_0 to calculate cosine for each document in the population, using vector space model with cosine similarity function, and sort the result in descending order (vectors are normalized).

$$Q_0: <C_{01}, C_{02}, \ldots C_{0n}>$$

   Select K-1 document based on the initial retrieval result and make them as purely relevant as possible.

   For a term that is in the K-1 documents, if it is also in Q_0, set its weight to the weight in Q_0. Then use the K-1 documents as additional queries and calculate cosine for each document in the collection again.

$$Q_0: <C_{01}, C_{02}, \ldots C_{0n}>_0$$
$$D_1: <C_{11}, C_{12}, \ldots C_{1n}>_1$$
$$\ldots$$
$$D_{k-1}: <C_{k-1,1}, C_{k-1,2}, \ldots C_{k-1,n}>_{k-1}$$
$$\text{----------}$$
$$\Sigma_1, \quad \Sigma_2, \quad \ldots \quad \Sigma\text{å}_n$$

Sum up cosines for each document and sort the final scores.

Since this procedure consolidates scores using sum cosine measure, we call it sum-cosine procedure. In this procedure, since additional queries are automatically generated using top K-1 documents of initial retrieval, we will call the top K-1 plus Q_0 the top K documents hereafter. This procedure is based on two assumptions: 1) Sum-cosine is a good consolidated measure; 2) Additional queries so generated are better representation of user's information need than the original query alone. In the following section, we explain why such assumptions are theoretically solid.

Theoretical Investigation

In this section, we will prove assumption 1, and show assumption 2 is generally reasonable.

Assumption 1

Here we prove why the sum cosine score is adequate as an integrated score for retrieved documents.

Suppose that a user submit a query Q_0, and the system decides to use D_1 and D_2, the first two articles returned in the list, as additional queries, we then have the following cosine matrix produced by the original query and additional queries.

	D_1	D_2	D_3	D_4	...	D_n
Q_0*	C_{01}	C_{02}	C_{03}	C_{04}	...	C_{0n}
D_1	C_{11}	C_{12}	C_{13}	C_{14}	...	C_{1n}
D_2	C_{21}	C_{22}	C_{23}	C_{24}	...	C_{2n}
Sum	Σ_1	Σ_2	Σ_3	Σ_4	...	Σ_5

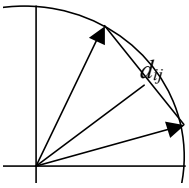
*Q_0 is original query, C_{ij} is the cosine of article i and j.

Analysis

Assume that the document set of D_1, D_2, and Q_0 is a set of the additional and the original query. Then what we need to show is that document D_3 through D_n should be ordered in a way that the one closer to the cluster should come first. The "closeness" of a document in D_1 through D_n to the cluster of (Q_0, D_1, D_2) can be measured in Euclidean distance of the document to the centroid of the cluster, i.e., the error square introduced, $d^2(D_k, Centroid)$. Then we need to prove that sum cosine is monotonically and inversely related to the error square.

Proof

Without loss of generality, suppose $\Sigma_3 > \Sigma_4$, we need to prove that D_3 introduces less error square to cluster (Q_0, D_1, D_2) than D_4 does. Let the distance between two normalized documents D_i, D_j, be d_{ij}. It is easy to show that cosine of vector D_i, D_j is a function of d_{ij}.



Angle and distance between two normalized vectors.

$$cos(D_i, D_j) = 1 - d_{ij}^2/2$$

Therefore,

$$cos(Q_0, D_3) + cos(D_1, D_3) + cos(D_2, D_3) = 3 - (d_{03}^2/2 + d_{13}^2/2 + d_{23}^2/2)$$

In the same way, we can show that

$$cos(Q_0, D_4) + cos(D_1, D_4) + cos(D_2, D_4) = 3 - (d_{04}^4/2 + d_{14}^4/2 + d_{24}^4/2)$$

and, $\Sigma_3 > \Sigma_4 \Rightarrow d_{04}^2 + d_{14}^2 + d_{24}^2 > d_{03}^2 + d_{13}^2 + d_{23}^2$

The above inequality implies that the sum of pair-wise square when D_4 is introduced to the cluster will be larger than if D_4 is not introduced. Because the sum of pair-wise square distance is the variance times number of observations in the cluster, therefore, D_4 introduces more error square. We proved that sum cosine is monotonically and inversely related to the error square introduced by adding a document. □

Also notice that how close a document is to the cluster is measured by how close it is to the centroid of the cluster (measured by error square). The centroid is an estimate of user's information need. This leads us to the assumption 2: how good can the centroid of top K represent user's information need?

Assumption 2

In assumption 2, we need to show that the top K documents so selected should be a better representation of user's information need.

User's information need in IR could be a rather behavioral or psychological issue. But in an experiment environment, it is usually assumed that the user's information need is clear and fixed, and thus documents can be clearly classified as relevant or not. (van Rijsbergen 1979, p113). In another word, the collection of relevant documents in corpus for each query is known and will satisfy user's information need.

Given these constrains, we can define user's information need as follows.

$$IN = \frac{1}{n}\sum_{R}\frac{D_i}{|D_i|}$$

Where *IN* is a vector for information need, *R* is the collection of relevant documents of size *n* in corpus, and D_i is the vector of a specific document. Given the definition of true user's information need, we can then measure the similarity of an estimated vector and the true one.

To investigate when and why top K can better approximate user's information need, let's further assume the top K are all relevant, then we can treat the top K as a sample from relevant documents. In practice, top K are not all relevant. The sample quality issue is treated separately in (Xu, 2001). Let's also follow the standard assumption in IR that terms in a document are independent, we can then analyze the property of a document vector by its elements.

Random sampling

Assume that the query itself is relevant, i.e., if the query is to be treated as a document, it will be classified as relevant by user. Also assume that the query is a relevant document randomly generated by the same underlying process that generates the relevant documents, then we have:

$$E(Q_0)=\boldsymbol{\mu}, \text{ and } \boldsymbol{\sigma}^2(Q_0)=\boldsymbol{\sigma}^2$$

Where $E(Q_0)$ is the expected value of original query, $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}^2$ are mean and variance of all relevant documents.

In the simplest case, if K documents are a random sample from the population of relevant documents, then it is obvious that the estimate of population mean from such sample will be more accurate than the original query regardless of the underlying term distribution. Mathematically, let **K** be the collection of top K documents, let $\overline{\mathbf{K}}$ be the average vector of top K document, it can be shown that:

$$E(\overline{\mathbf{K}})= \boldsymbol{\mu}, \text{ and } \boldsymbol{\sigma}^2(\overline{\mathbf{K}})= \boldsymbol{\sigma}^2/K$$

Non-random sampling

The problem is that the top K generating process is not a random sampling process and the term distribution is usually a mixture of two Poisson distributions. Since terms are independent, let's distinguish terms that appear in Q_0 and those do not. For terms in Q_0, if top K evenly distribute on both sides of Q_0, the expected value of sample average is rarely the population mean, as illustrated below.

Fortunately, for terms not in the original query, since they are independent of those in original query, the top K can still be viewed as a random sample. If the original query is relatively small in size compared to the collection of all terms in the top K, we can still argue that the top K average is a better estimate in most dimensions, while in a small portion of dimensions it produces uncertain estimate. If the benefit gained from terms not in the original query out weights the loss from those in it, we can still expect a better similarity measure between the top K average and the true mean.

To avoid the biased estimate of terms in Q_0, we don't use the simple average of top K as the estimate of information need. Instead, we treat terms in Q_0 and those don't differently. For those in Q_0, we keep its original weight in Q_0; for those not, we use the sample average. In this way, we gain accuracy from terms not in Q_0 and lose no accuracy for terms in Q_0. We set the top K centroid to this estimate.

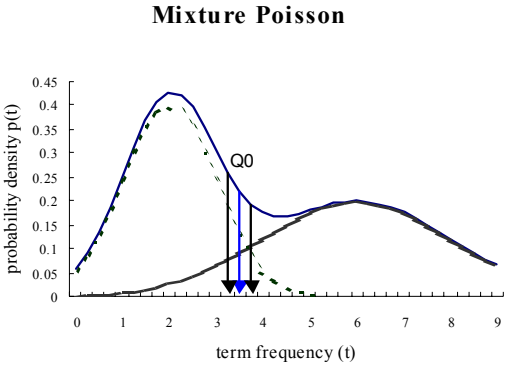


Figure 2. Asymmetric Term Distribution

It should also be seen that the sum-cosine procedure is essentially automatic relevance feedback which estimates information need with a special sampling method rather than adaptively revised query.

Sources of error and selection of top K
So far we show that the two assumptions underlying our procedure are solid. It also reveals that the major source of error in the estimate is irrelevant documents in the top K. To mitigate this source of error, top K must be purified so that only relevant documents are kept. The detailed purification methods are discussed in (Xu, 2001). Here we give a coarse heuristic rule to select top K, which is based on the observation of the similarity data. We found that the cosine series obtained after step one is an "exponentially" decreasing series. The similarity difference between two consecutive relevant documents are usually fluctuating heavily, while for irrelevant documents, it is consistent and small. We thus set our rule that if the standard deviation of five consecutive differences is below 20% of the highest standard deviation of differences seen so far, the documents involved and thereafter are irrelevant.

Relevance feedback revisited
Compared with automatic relevance feedback, which adjusts the estimate of user's information need with a weighted sum of last previous estimate and new documents, this procedure tries to construct an unbiased estimate from a relevant sample. This is the key difference between the two. As will be shown later, our experiments show that the difference is significant.

Empirical Tests

Description of Data Sets

We test our procedure on four data sets which have been used in other IR studies. A short summary of these data sets is given in table 1

For each query, the data set also provides a corresponding list of relevant documents. We thus have a "hard" standard to judge the performance of our search mechanism. Queries are in natural language form.

Table 1. Data Sets

Data sets	Description	No. of docs	No. of query tested
Time	Time Magazine full text articles in 1963	425	60
Medline	Medical literature abstracts	1033	30
Cranfield	Mechanical literature abstracts	1400	225
CISI	Computer Information Science literature abstracts	1460	112

Experimental Setting

For each document, we do the regular preprocessing, including stop words deletion and Porter's stemming. When converting a document to a vector, we keep all the words in it, because the abstracts are usually short. TFIDF weighting scheme is used with normalization. The top K is selected using heuristic rule described in section 4.

There are a few goals that our experiments want to achieve:

- The performance of the sum-cosine procedure and its benchmarks
- How well does the top K capture user's information need?

To achieve the first goal, there are three benchmarks tested. The vector space model benchmark uses standard vector space model with cosine similarity measure. The top 10 relevance feedback benchmark uses automatic relevance feedback with only positive feedback from top 10 articles. The cut-off feedback uses the heuristic rule to pick top K documents. We use the same dynamically picked top K documents in sum-cosine procedure.

To achieve the second goal, we compare the similarity between the top K centroid and the actual information need (mean of all relevant documents), and the similarity between original query and actual information need.

Experiment Results

Average Precision
Here we report the precision-recall measurements and average precision of the sum-cosine procedure and benchmarks. The average precision is the average precision over all recall levels, which serves as a rough measurement of the overall performance of an algorithm.

Table 2. Average Precision

Datasets	Std. Vector space	Top 10 RF	Cut-off RF	Sum cosine
Time	77%	74%	79%	84%
Medline	56%	62%	62%	66%
Cranfield	33%	34%	36%	38%
CISI	26%	26%	26%	28%

The experiment results show that sum-cosine procedure performs significantly better than all benchmarks. To see that sum-cosine is superior to relevance feedback, we should notice even when they use the same top K documents as the core to estimate user's information need, the performance is still significantly different.

Estimate of user's information need
To test how the selected top K can better estimate user's information need, we generate the centroid of top K for each query, and test its similarity with the true information need. We show the following statistics.

Table 3. Similarity of Top K and Initial Query

Datasets	Top K (stdev)	Initial Query (stdev)	p-value	% of Better Estimate
Time	0.19 (0.16)	0.0007 (0.0006)	3.9E-17	95%
Medline	0.12 (0.045)	0.0075 (0.0069)	2.4E-16	100%
Cranfield	0.09 (0.035)	0.0070 (0.0064)	2.4E-53	92%
CISI	0.09 (0.0065)	0.0065 (0.0057)	6.8E-38	100%

It should be easy to see that the average cosine of top K centroids is significantly better than that of the original queries. As a mater of fact, in Medline and CISI, for all queries, cosines of the top K are better than that of the original queries. In other two data sets, more than 90% of queries have cosine of top K better than that of original query.

Conclusion

Our research proposed a new meta-search procedure which uses top K documents of initial retrieval as additional queries and integrates multiple retrieval runs with consolidated cosine score. We show that the improvement in performance is gained through better estimation of user's information need and ordering documents in according to least error square criterion. This is supported by both theoretical analysis and empirical evidence. As stated in section 4, the major source of error in this procedure is the existence of irrelevant documents in top K. This could be a limitation of this study. It is projected that better purification methods for top K in future research will further improve performance.

Reference

Efthimiadis, E.N. "Interactive Query Expansion: a user-based evaluation in a relevance feedback environment." *Journal of the American Society for Information Science* (51:11), 2000, pp. 989-1003.

Ng, K.B. and Kantor, P. "Predicting the Effectiveness of Naive Data Fusion on the Basis of System Characteristics". *Journal of American Society for Information Science* (51:13), 2000, pp. 1177-1189

Belkin, N.J., Cool, C., Croft, W.B. and Callan, J.P. "The effect of multiple query representations on information retrieval performance," *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1993, pp. 339-346

Bartell, B.T., Cottrell, G.W. and Belew R.K. "Automatic combination of multiple ranked retrieval systems," *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* 1994, pp. 173-181.

Fox, E.A. and Shaw, J.A. "Combination of multiple searches," *Proceedings of the Third Text Retrieval Conference* (*TREC-3*), National Institute of Standards and Technology Special Publication 500-215, 1994, pp. 243-252.

Lee, H.J. "Analyses of multiple evidence combination," *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1994, pp. 267-276.

Katzer, J., McGill, M.M., Tessier, J.A., Frakes, W., and Dasgupta, P. "A study of the overlap among document representations," *Information Technology: Research and development* (1:2), 1982, pp. 261-274.

Rocchio, J.J. Jr., "Relevance Feedback in Information Retrieval," *Scientific Rpt. ISR-9, Section 23, Harvard Comp. Lab.,* Cambridge, MA, Aug. 1965.

Salton, G., and Buckley, C. "Optimization of Relevance Feedback Weights," *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* Seattle, USA, 1995, pp. 351-357.

Xu, Y., "Data fusion with multiple queries in single information retrieval scheme," *Technical repor*, Department of Quantitative Methods, Syracuse University, 2001.